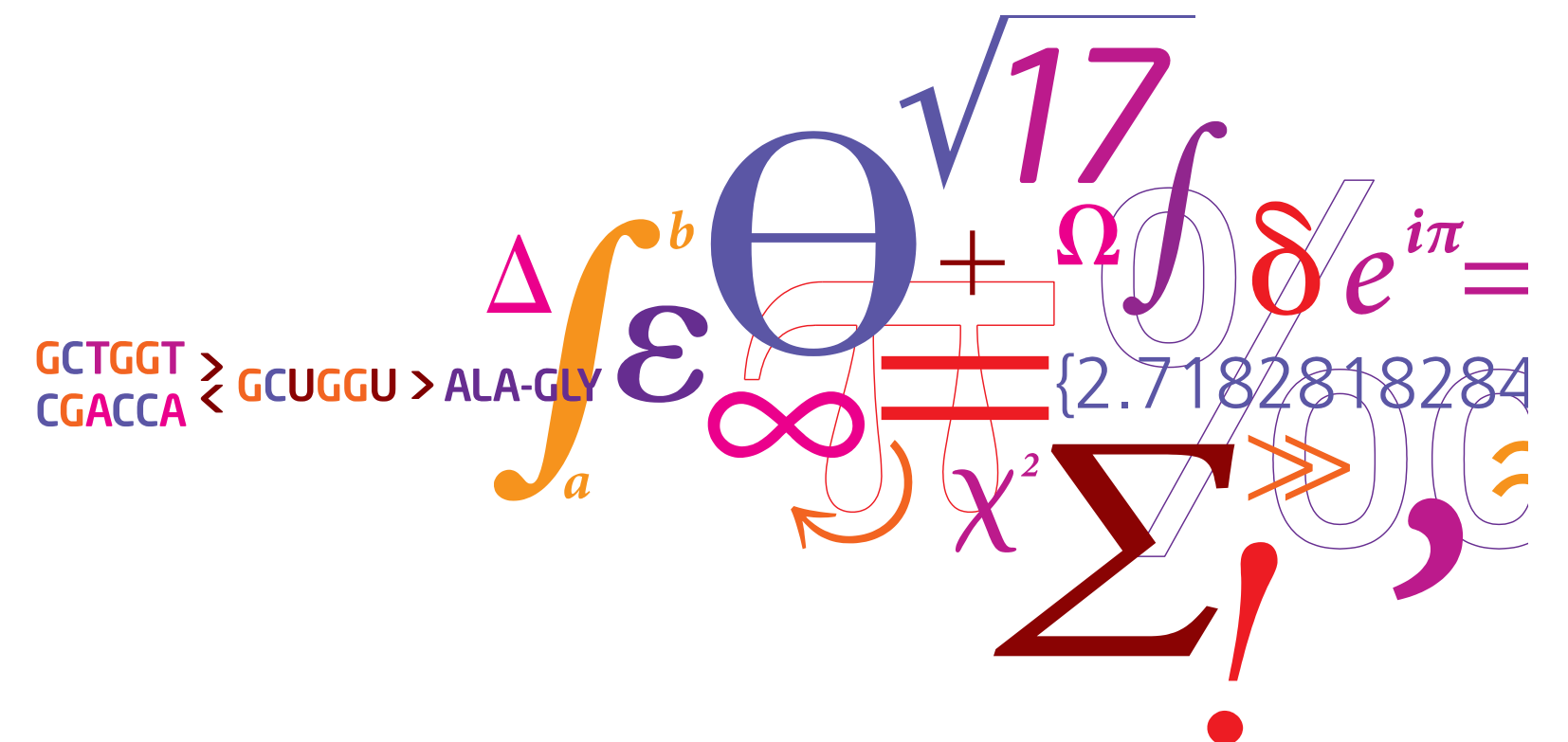


Introduction to Next Generation Sequencing (NGS)

Simon Rasmussen
Assistant Professor
Center for Biological Sequence analysis
Technical University of Denmark
2012



Today

- 9.00-9.45: Introduction to NGS, How it works
- 10.00-10.30: Data basics - what does the data look like?
- 10.30-11.00: *de novo* assembly exercise
- 11.15-12.00: Alignment of reads
- 13.00-13.30: Introduction to exercise (variations, alignment processing, genotyping)
- 13.30-16.30: Afternoon exercise

DNA sequencing

**Reading the order
of bases in DNA
fragments**

Why NGS?

Transforming how we are doing
biological science (and bioinformatics)

by

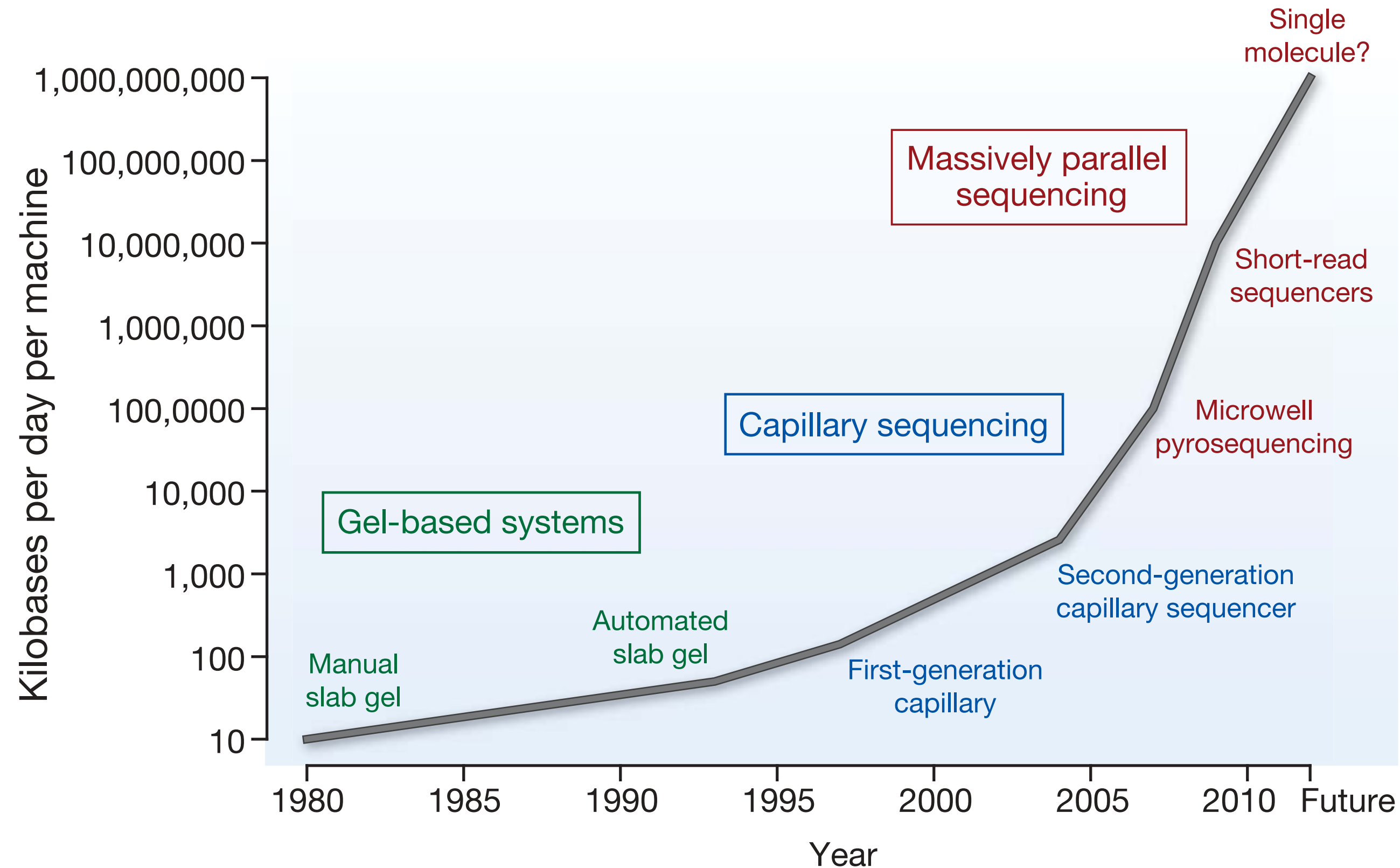
allowing experiments that could not
have been done before, and perform
experiments *much* faster

How ?



by producing **massive** amounts of sequence data, really **fast**

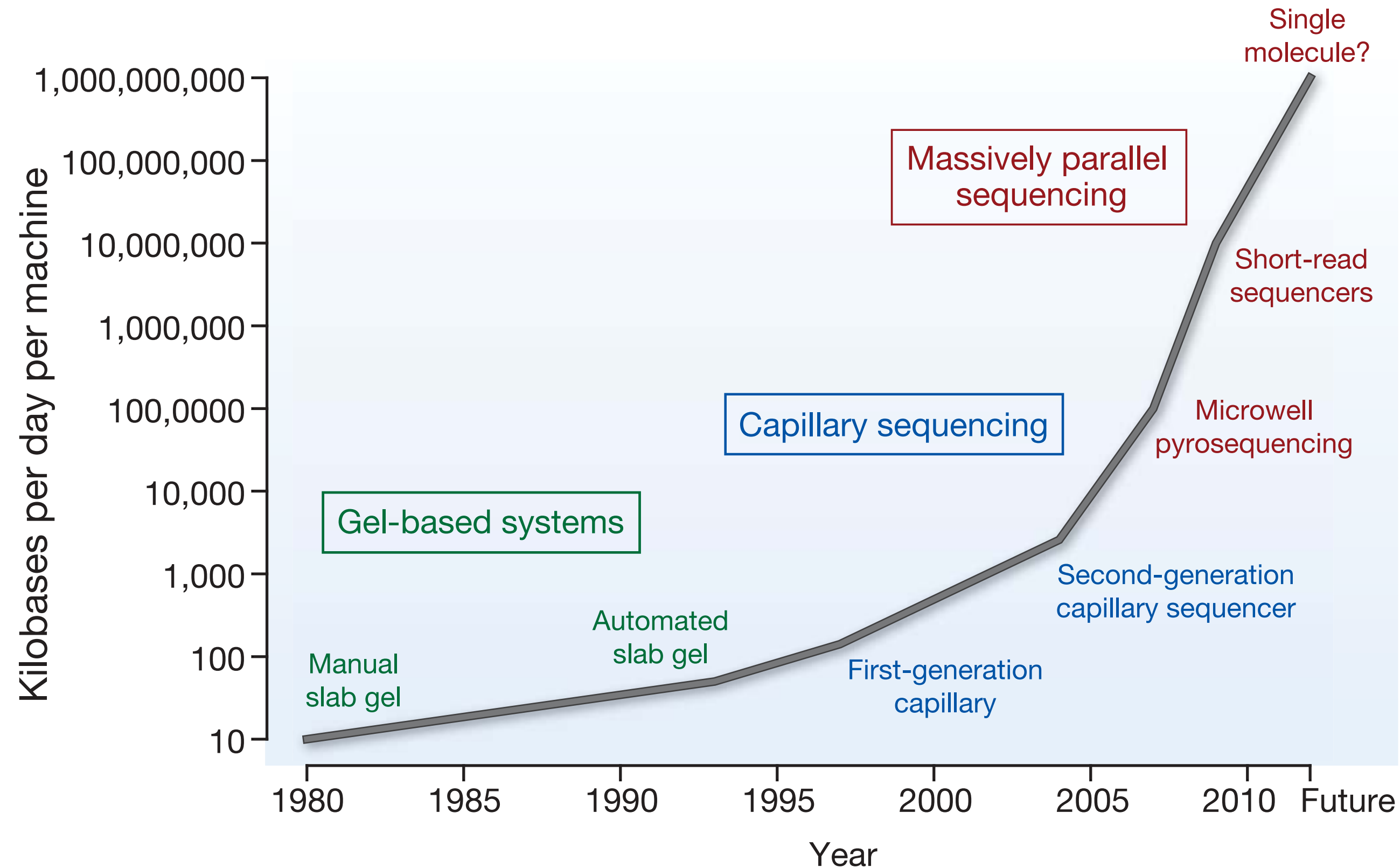
1st generation to NGS



1977 - Sanger
Chain-termination
method

Stratton et al., Nature 2009

1st generation to NGS

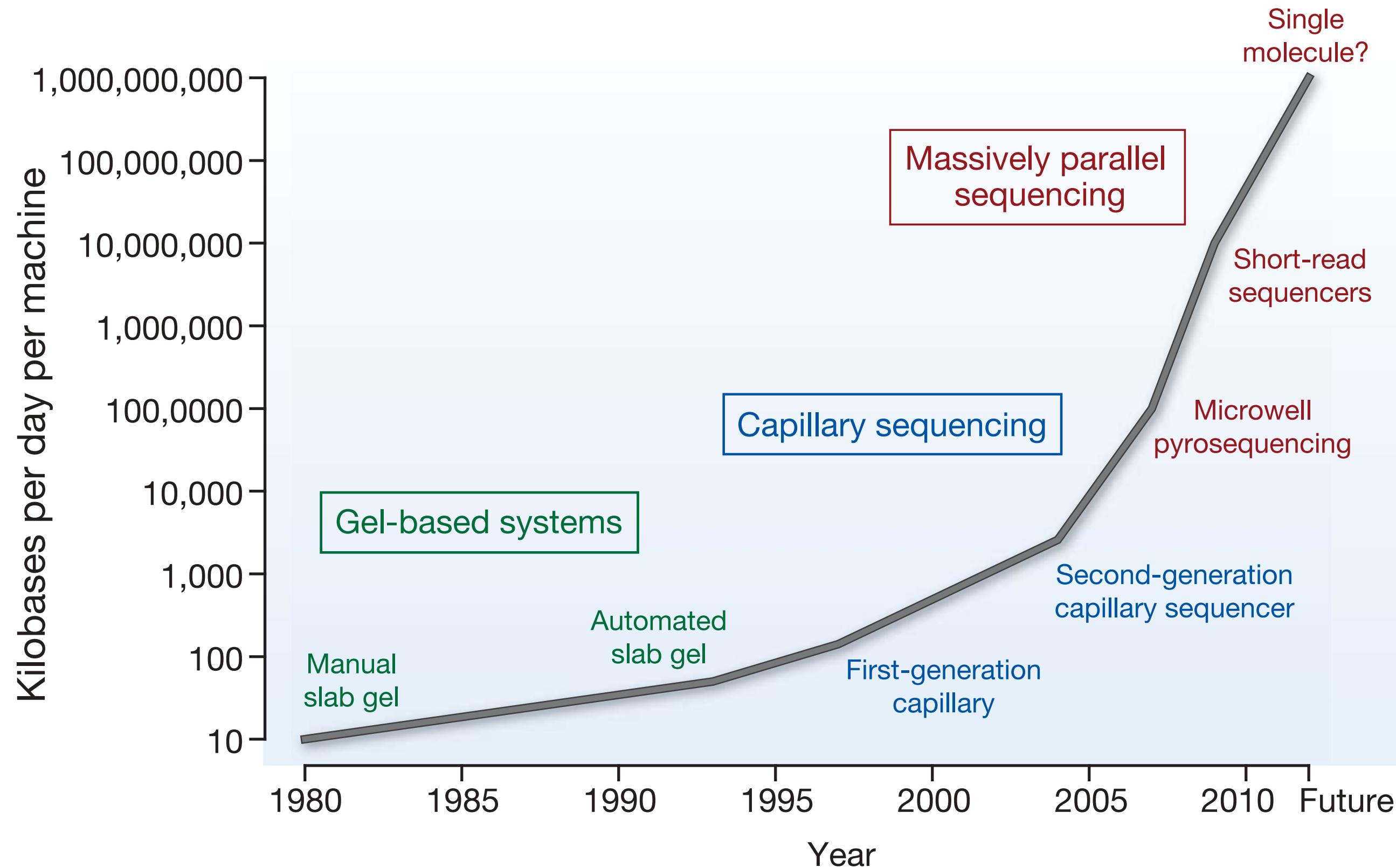


1977 - Sanger
Chain-termination
method

Human genome

Stratton et al., Nature 2009

1st generation to NGS



1977 - Sanger
Chain-termination
method

Human genome

Illumina
Solid

454

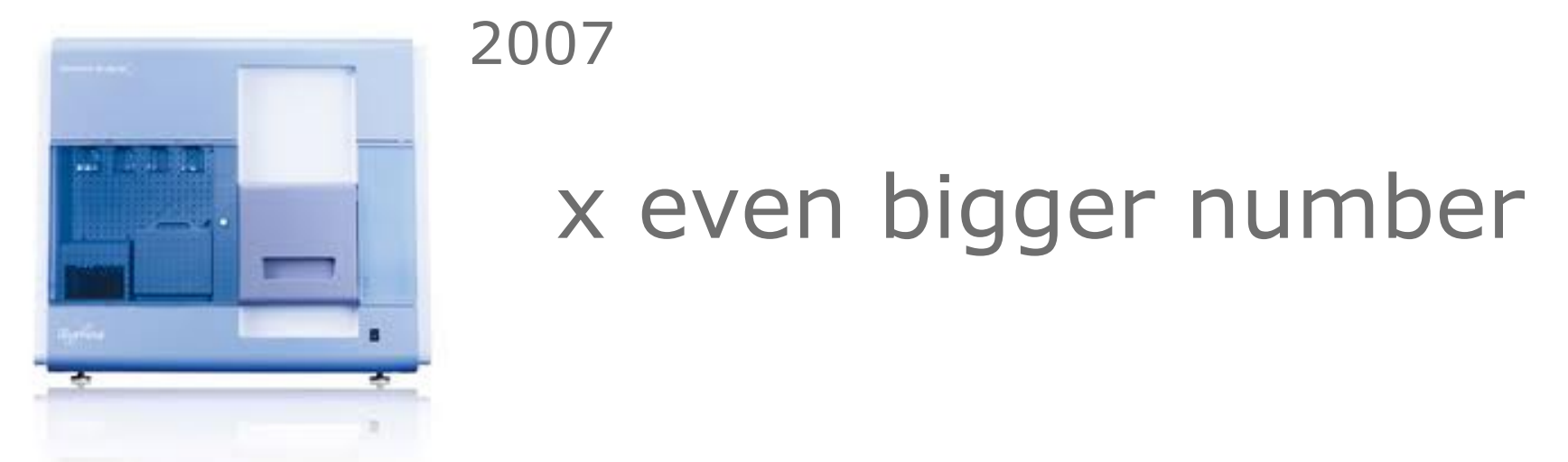
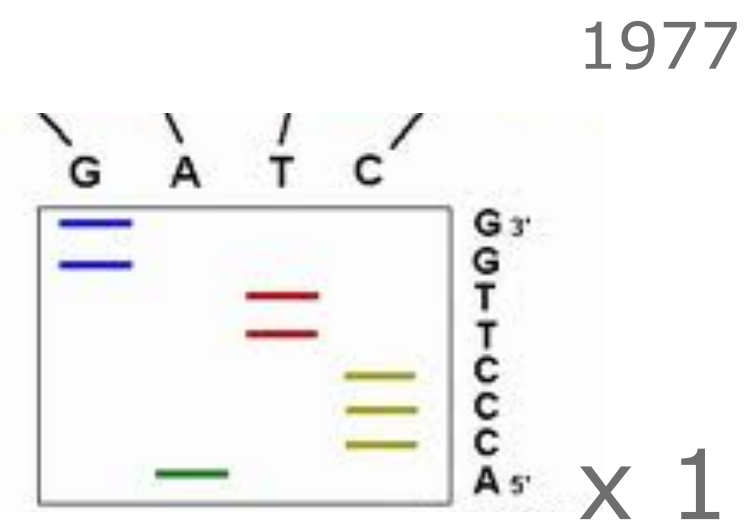
Ion Torrent

Pacific
Biosciences

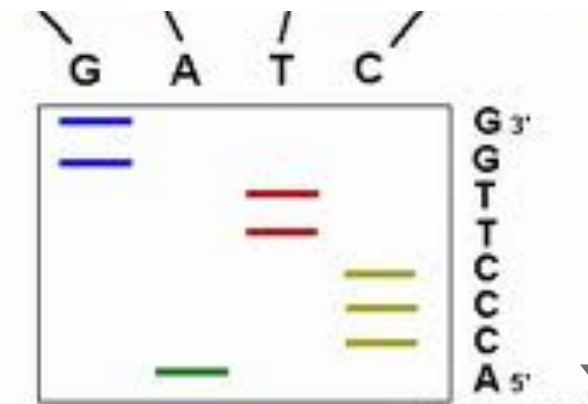
Oxford
Nanopore

Stratton et al., Nature 2009

Read throughput



Read throughput



1977

x 1

I - 384



1998

x 384



2006

x very big number



2007

x even bigger number



2008

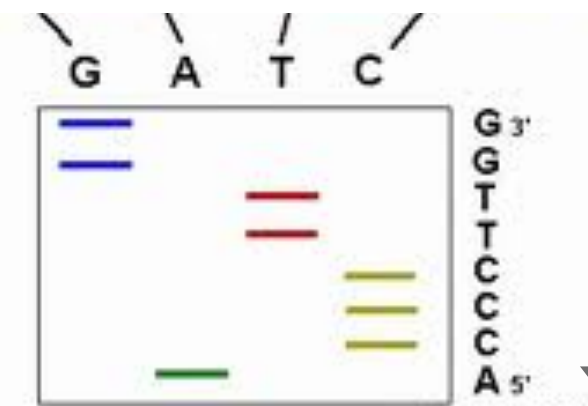
x gigantic number



2011

x big number

Read throughput



1977

x 1

1 - 384



1998

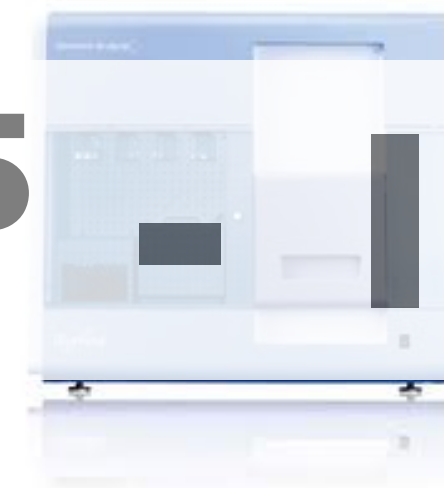
x 384



2006

x very big number

10⁵



2007

10⁹

x even bigger number



2008

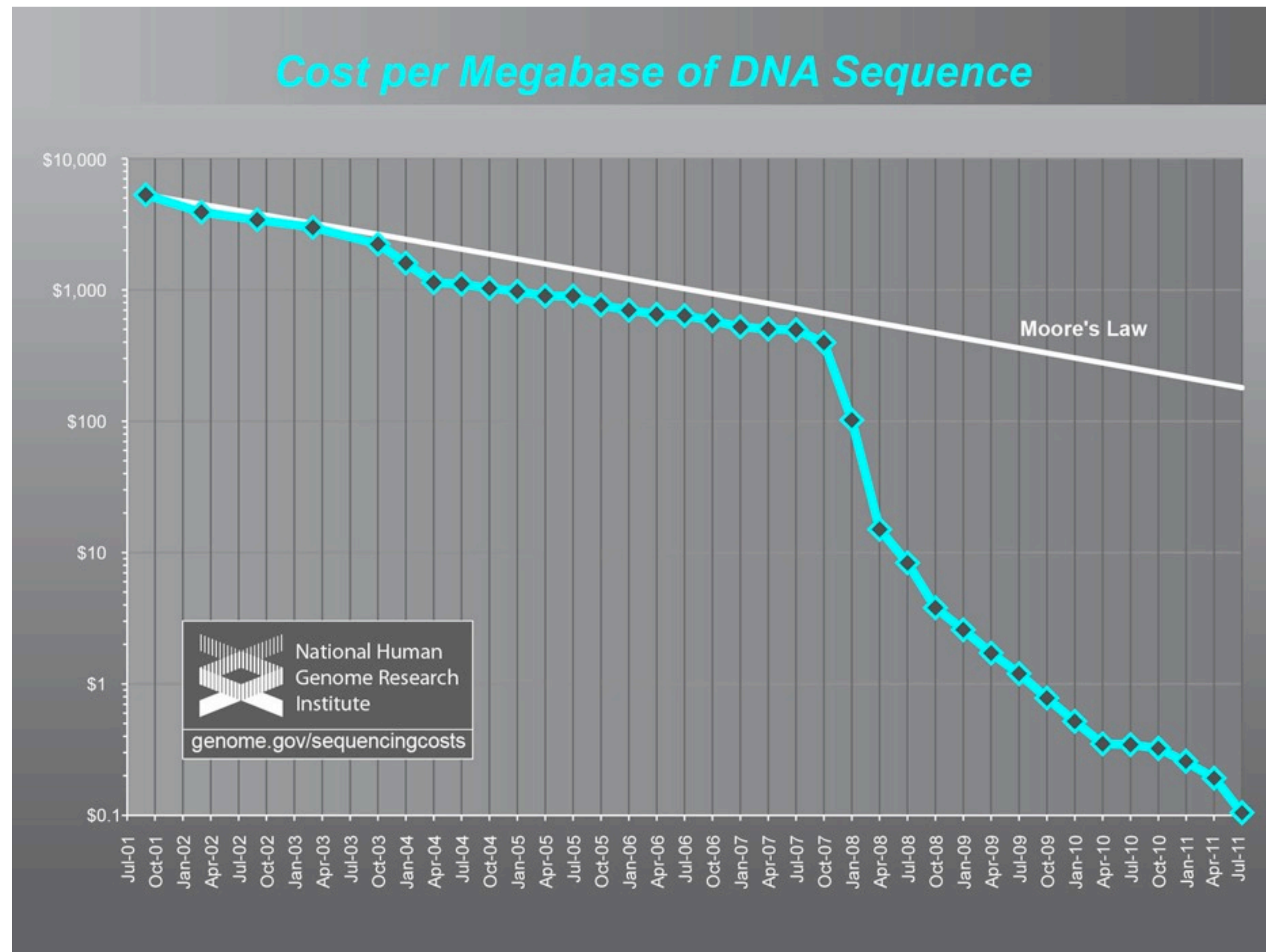
x gigantic number



2011

x big number

Sequencing costs



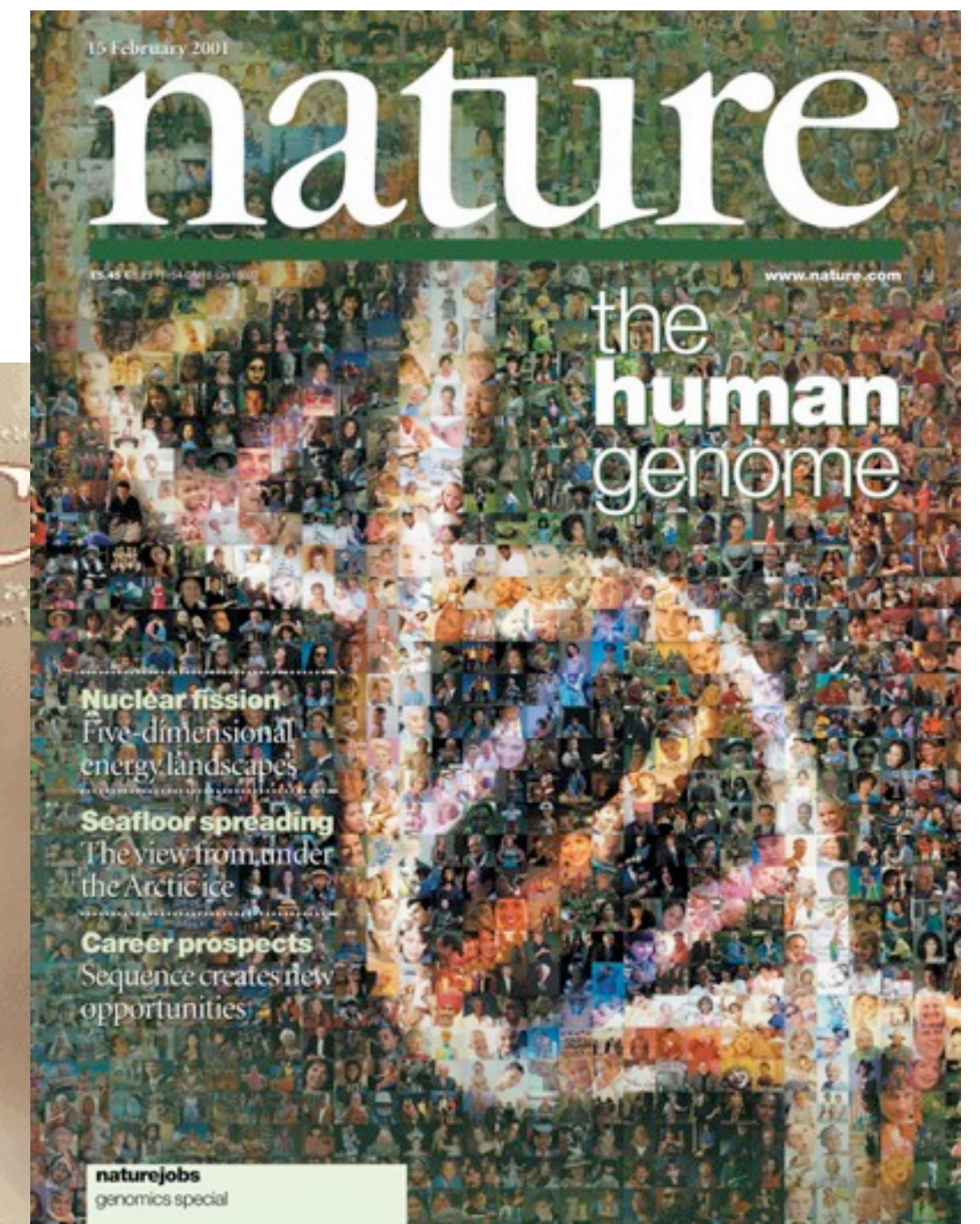
Drop in costs is faster
than Moore's Law

*(Computer power doubles
every 2 years)*

Human sequencing

- First draft genome of human in 2001, final 2004
- Estimated costs \$3 billion, time 13 years
- Today:
- Illumina: 1 week, 4000\$
- Exome: 6 weeks*, \$998
- Towards 1000\$ genome?

* Real-time, not machine-time



Storage and analysis



Highest cost is (almost) not the sequencing
but
storage and analysis

A standard human (30-40x) whole-genome sequencing exp. would create 100 Gb of data

Storage and analysis



Highest cost is (almost) not the sequencing
but
storage and analysis

A standard human (30-40x) whole-genome sequencing exp. would create 100 Gb of data

[BGI](#), based in China, is the world's largest genomics research institute, with 167 DNA sequencers producing the equivalent of [2-4,000 human genomes a day](#).

The X Genomes projects

- 1000 genomes project:
 - Catalog of human genetic variation, including SNPs and structural variants, and their haplotype contexts
 - Sequence 2500 unidentified people from about 25 populations around the world
- 10.000 microbial genomes project, Earth Microbiome project, Cancer genome project, Plants and animals, ...

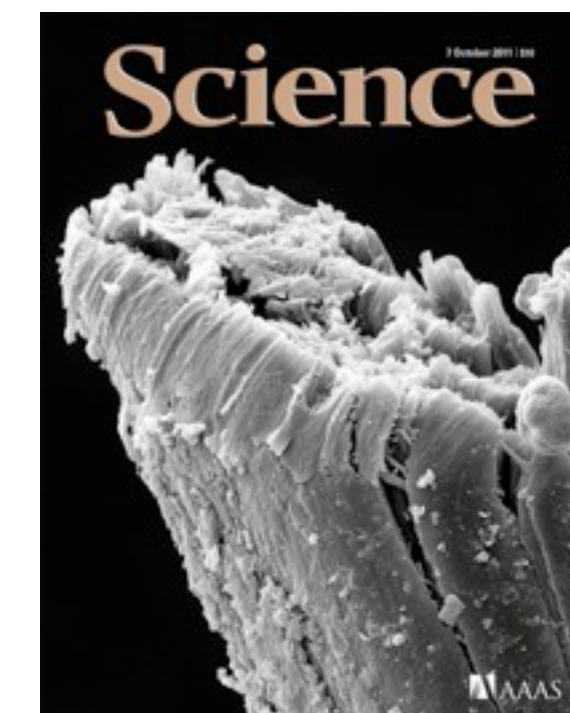
NGS & Bioinformatics

- Extreme data size causes problems
 - Just transferring and storing the data
 - Standard comparisons fail (N^2)
 - Standard tools can not be used
 - Think in fast and parallel programs



What can we use it for?

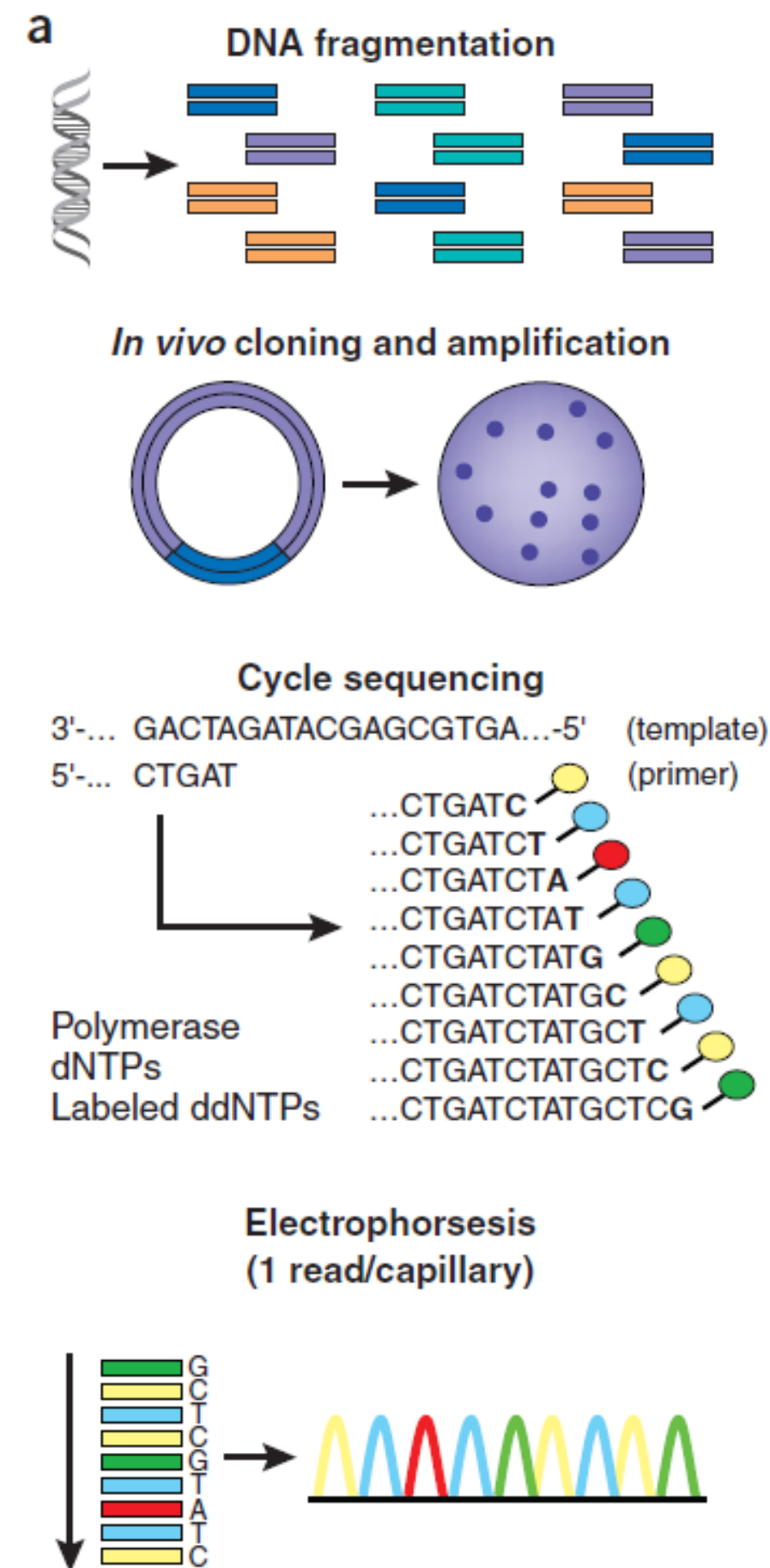
- Whole genome re-sequencing
- Ancient genomes
- Metagenomics
- Cancer genomics
- Exome sequencing (targeted)
- RNA sequencing
- Chip-seq
- Genomic Epidemiology
- *anything with DNA*



How it works?



First generation: Sanger (dye)



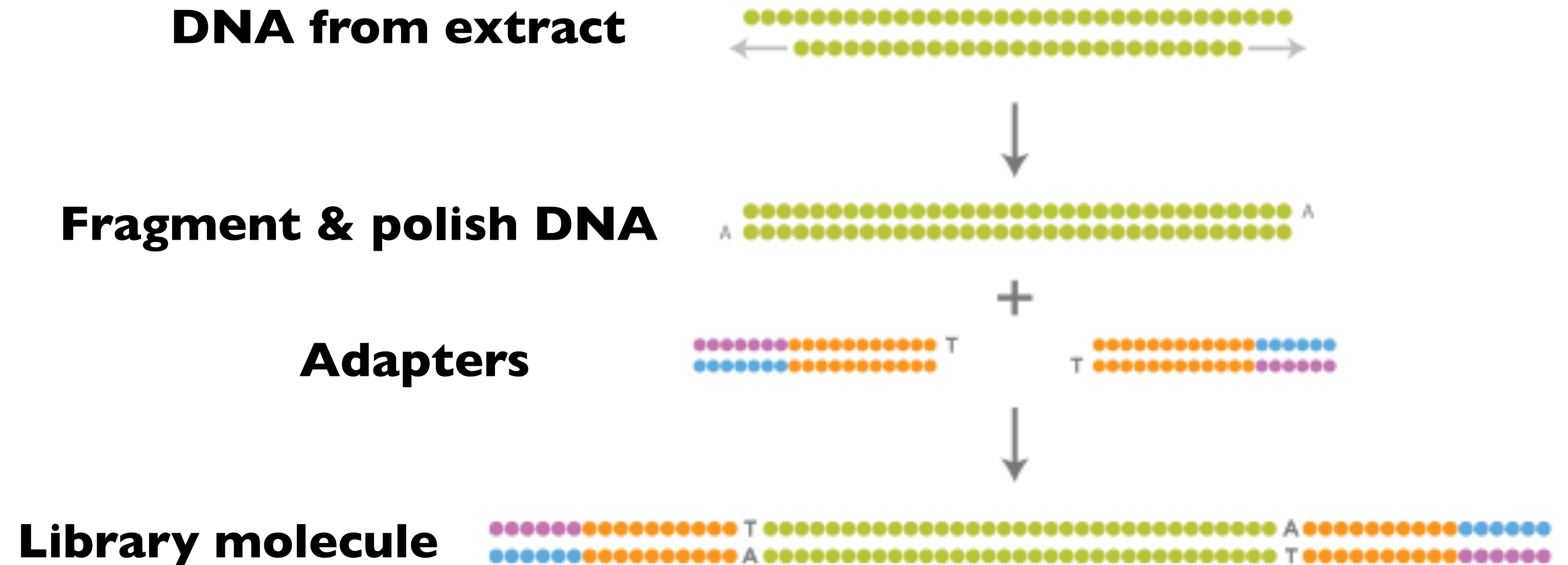
- Fragment DNA
- Clone into plasmid and amplify
- Sequence using dNTP + labelled ddNTPs (stops reaction)
- Run capillary electrophoresis and “read” DNA code
- Low output, long reads (~300-1000 nt), high quality

2nd generation

- 1. Create library molecule**
- 2. Amplification (PCR)**
- 3. Massive parallel sequencing**

2nd generation

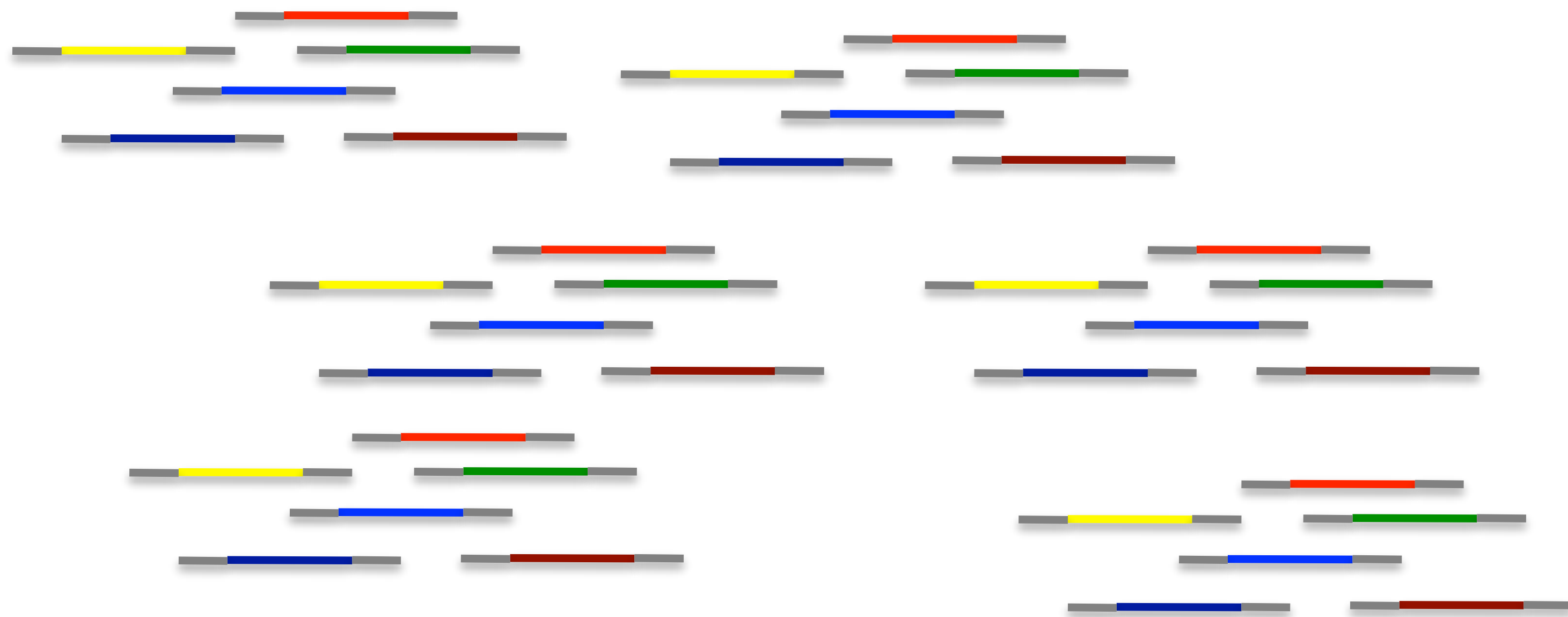
1. **Create library molecule**
2. **Amplification (PCR)**
3. **Massive parallel sequencing**



2nd generation

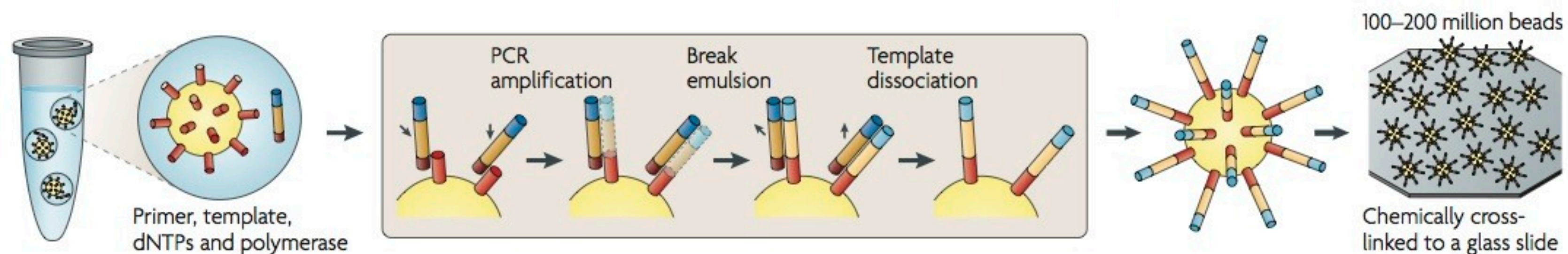
1. **Create library molecule**
2. **Amplification (PCR)**
3. **Massive parallel sequencing**

Library

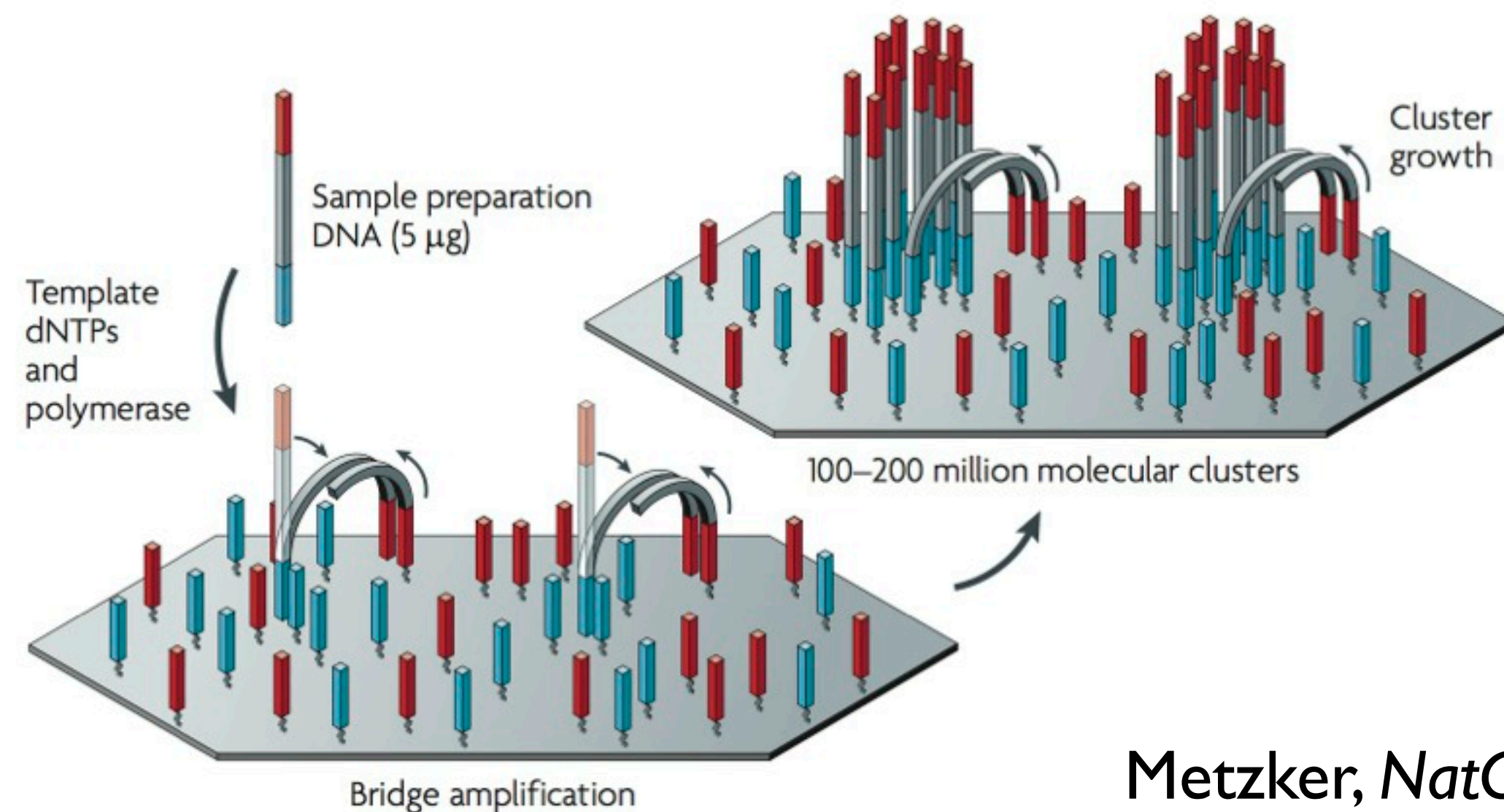


Amplification and immobilization

Emulsion PCR (454, Solid, IonTorrent): Water, oil, beads, one DNA template/droplet



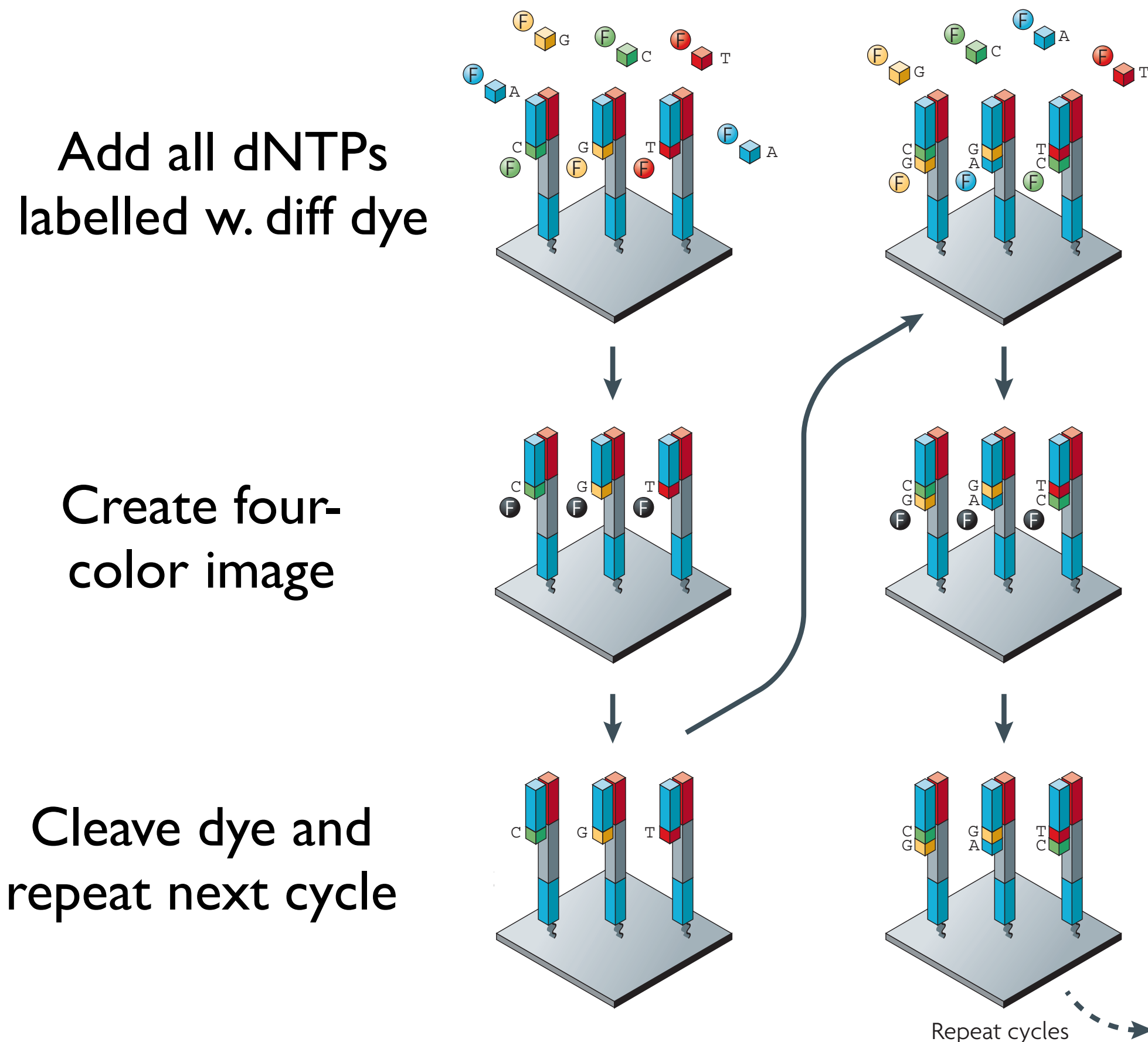
Bridge PCR (Illumina):
One DNA template/cluster,
primers on surface,
grow by bridging primers



Metzker, *NatGen Rev.* 2010

Fluorescence detection

Illumina - Cyclic reversible termination



454 - Pyrosequencing

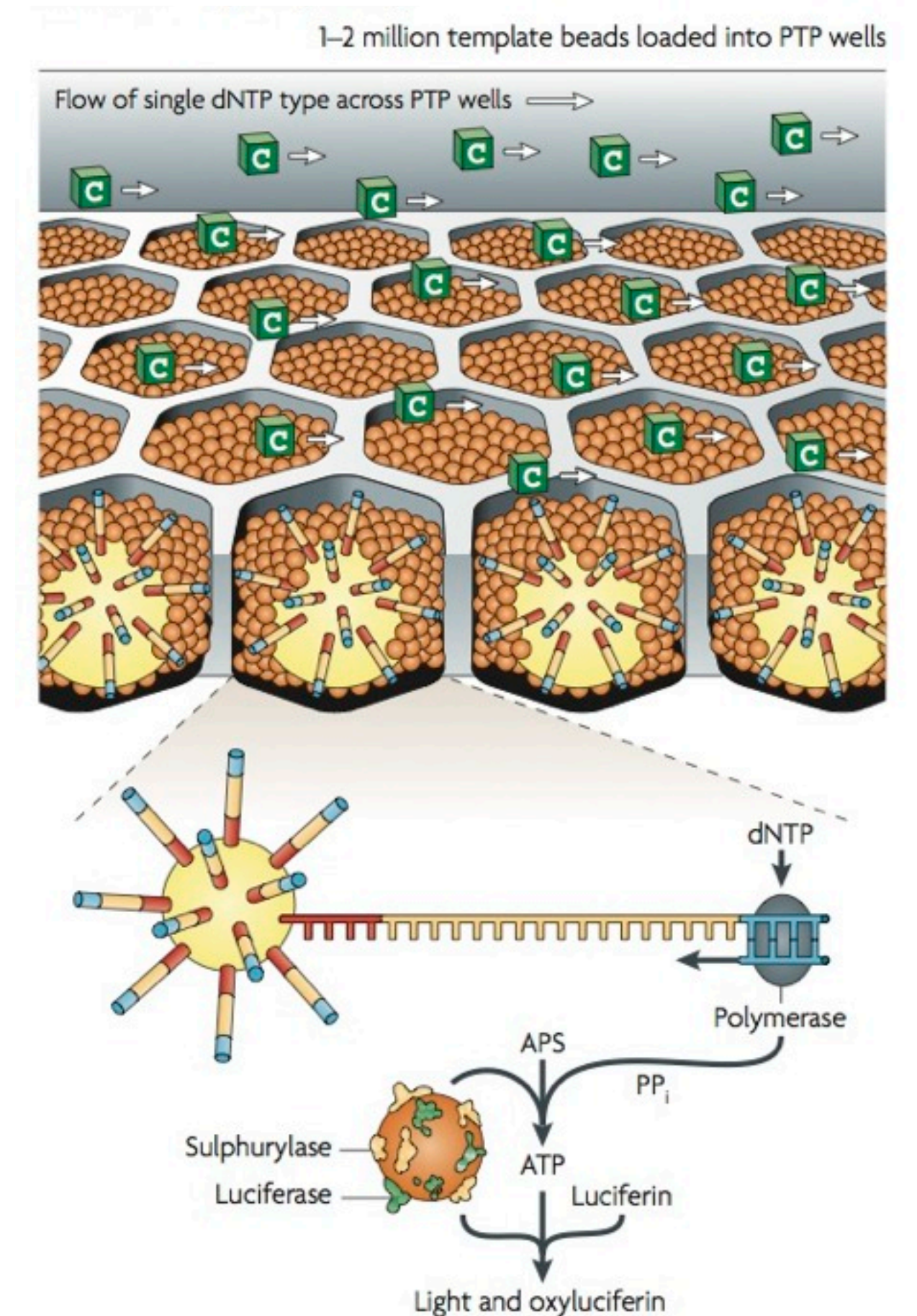
Load template beads into wells

Flow one dNTP across wells

Polymerase incorporates nucleotide

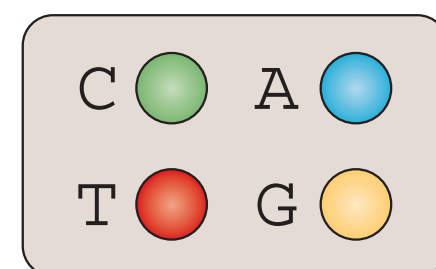
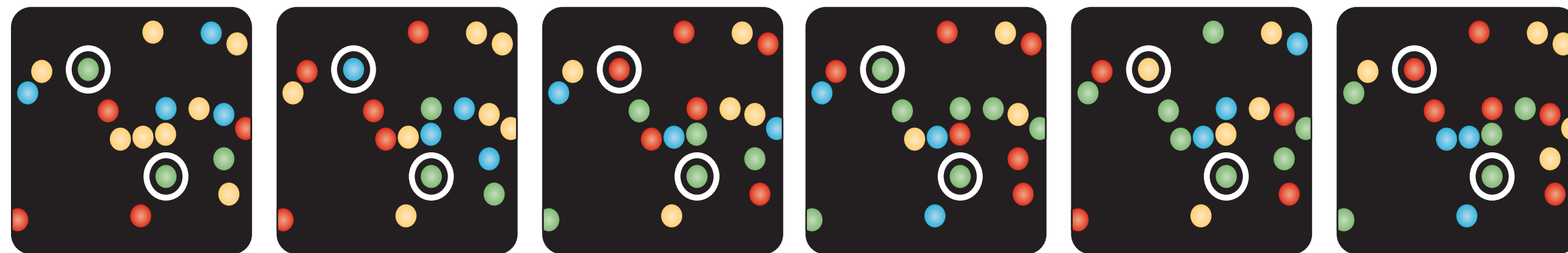
Release of PP_i leads to light

Imaging, next dNTP



Metzker, *NatGen Rev.* 2010

2G: Imaging handout



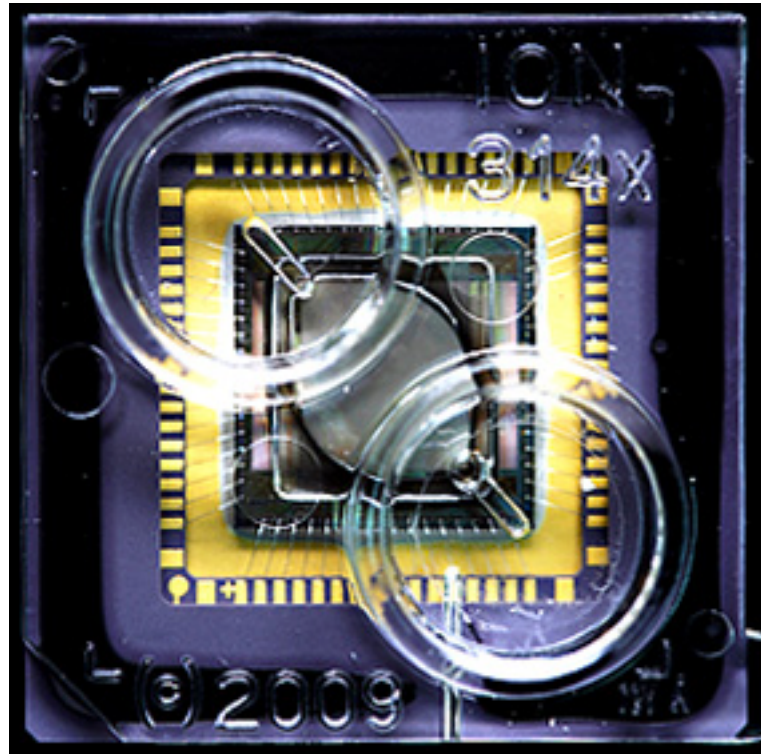
Illumina 1: _____

Illumina 2: _____



454: _____

2.5G: Ion Torrent



[IonTorrent video](#)

- Based on semiconductors, ie. no fluorescence
- Release of hydrogen when a nucl. is incorporated is measured by ph-meter
- Small machine, low price pr. run

3rd generation

No amplification (PCR introduces bias!)

Simple sample preparation



Helicos



Pacific Biosciences



Oxford Nanopore

Platform	3730XL	454 FLX	454 GS JR	HiSeq 2000	MiSeq	SOLiD 5500	IonTorrent	PacBio RS
Method of amplification	Clonal plasmid amplification	emRCR on beads	emRCR on beads	Bridge PCR amplification	Bridge PCR amplification	emPCR on bead	emPCR on bead	None
Chemistry	Chain termination	Synthesis (Pyro-sequencing)	Synthesis (Pyro-sequencing)	Synthesis (Reversible termination)	Synthesis (Reversible termination)	Ligation (dual-base encoding)	Synthesis (H ⁺ detection)	Synthesis
Instrument Cost	\$376k	\$500k	\$108k	\$690k	\$125k	\$595k	\$67.5k	\$695k
Yield per Run	60 kb	900 Mb	50 Mb	600 Gb	1 Gb	155 Gb	1 Gb	20-80 Mb
Read Length (bases)	650	750	400	100	150	75 + 35	200 (318 chip)	<1,800 - >5,000
Reagent Cost (library + run)	\$96	\$6 200	\$1 100	\$23 610	\$1 035	\$10 503	\$925	\$272
Cost per Mb	\$1600	\$7	\$22	\$0.039	\$1	\$0.068	\$0.93	\$3.4-13.6
Primary error & error rate	Substitution 0.1-1 %	Indel 1%	Indel 1%	Substitution >0.1%	Substitution >0.1%	indel >0.01%	Indel ~1%	Indel ~15%
Primary Advantage	Low cost for small study	Long read length	Long read length	Most output at lowest cost	Easy workflow & fast run	Each lane can be run independently & ability to rescue failed cycle	Fast run, low cost, and trajectory to longer read	Longest read length, single molecule real-time seq
Primary Disadvantage	High cost for large study	Unreliable for homopolymer region; High cost NGS	High cost per Mb	High capital cost & computation need	Few reads & higher cost per Mb	Relatively short read, more gap in assemblies	Unreliable for long homopolymer region	High error rates, Low output, expensive

Platform	3730XL	454 FLX	454 GS JR	HiSeq 2000	MiSeq	SOLiD 5500	IonTorrent	PacBio RS
Method of amplification	Clonal plasmid amplification	emRCR on beads	emRCR on beads	Bridge PCR amplification	Bridge PCR amplification	emPCR on bead	emPCR on bead	None
Chemistry	Chain termination	Synthesis (Pyro-sequencing)	Synthesis (Pyro-sequencing)	Synthesis (Reversible termination)	Synthesis (Reversible termination)	Ligation (dual-base encoding)	Synthesis (H ⁺ detection)	Synthesis
Instrument Cost	\$376k	\$500k	\$108k	\$690k	\$125k	\$595k	\$67.5k	\$695k
Yield per Run	60 kb	900 Mb	50 Mb	600 Gb	1 Gb	155 Gb	1 Gb	20-80 Mb
Read Length (bases)	650	750	400	100	150	75 + 35	200 (318 chip)	<1,800 - >5,000
Reagent Cost (library + run)	\$96	\$6 200	\$1 100	\$23 610	\$1 035	\$10 503	\$925	\$272
Cost per Mb	\$1600	\$7	\$22	\$0.039	\$1	\$0.068	\$0.93	\$3.4-13.6
Primary error & error rate	Substitution 0.1-1 %	Indel 1%	Indel 1%	Substitution >0.1%	Substitution >0.1%	indel >0.01%	Indel ~1%	Indel ~15%
Primary Advantage	Low cost for small study	Long read length	Long read length	Most output at lowest cost	Easy workflow & fast run	Each lane can be run independently & ability to rescue failed cycle	Fast run, low cost, and trajectory to longer read	Longest read length, single molecule real-time seq
Primary Disadvantage	High cost for large study	Unreliable for homopolymer region; High cost NGS	High cost per Mb	High capital cost & computation need	Few reads & higher cost per Mb	Relatively short read, more gap in assemblies	Unreliable for long homopolymer region	High error rates, Low output, expensive